Secure Data Capture from Digital Sources

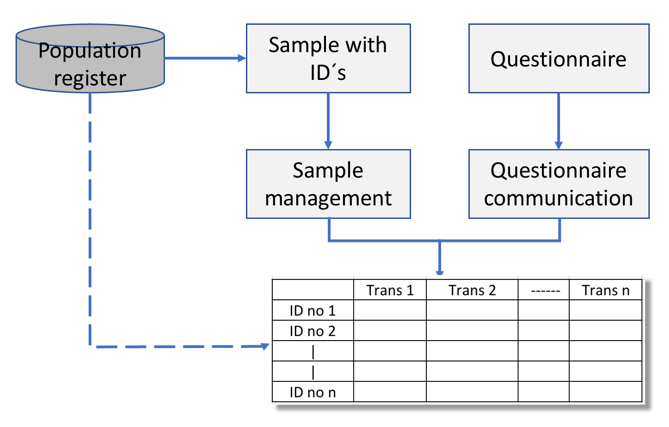
**Keywords:** Data collection, digital sources, privacy measures, data protection, management issues

1. **Introduction**

When we ask survey participants about actions or events in a questionnaire survey, we ask them to reproduce conditions that we could have observed if we had been present. For this type of questions, the digital footprints that the activities leave behind are a more direct source of data than the respondents' reproduction. The same applies when we ask about other factual factors such as occupation and income in personal surveys and production and turnover in business surveys. If we are able to replace this type of questions with digital sources, we get more precise results. In addition, we remove the workload for the respondents we might otherwise ask.

This is what we aimed for in a project where we, in an effort to replace the most burdensome part of consumer expenditure diaries with data from digital sources, tried to link receipts from grocery chains to debit information from bank accounts [1]. Even if this project, hereafter named the transaction project, was a success in that we were able to link purchases made in grocery stores to bank accounts, and from there to households, the project also illustrates how a traditional data collection approach leads to privacy concerns which complicate the realisation of the potentials of this kind of data. As an alternative to the traditional approach, this NTTS paper will present a compartmented data collection and processing system, which can greatly reduce the identification risk throughout the statistical value chain.

1. **From unit-based to behaviour-based data capture**

Traditional data collection is based on identifiable units; preferably personal id´s or organisation numbers. Basically, the procedure is as outlined in Figure 1.

**Figure 1. Unit-based data collection**

Security risks both occur because the identifications are stored, and because the results are gathered in a data matrix where id numbers are linked to the new information collected in the questionnaire. From this unit-by-variable matrix, statistical tables are generated.

Data capture and statistics production based on digital sources neither can, nor should be carried out in this traditional way. Unit identifications may either be absent or difficult to access because they are protected by privacy agreements or are the subject of commercial interests. The more personal information that is collected, the more difficult it becomes to maintain confidence that the data is not being misused. If we could produce statistics without linking the collected data to identifiable units, it would both be easier and less controversial to use digital sources for statistical purposes. This will open opportunities for better, quicker and richer statistical products.

This is what we aim for applying the compartmented system suggested in this paper. The essential feature of this system is that the groups of units and the statistical indicators that go into the statistical tables are produced separately and only linked together on an aggregated level. Consequently, the statistical table is produced without creating the full data matrix along the way. Because the collected data is never linked to identifiable units, there is no need for informed consent.

In the first compartment, named the Extraction Compartment, three kinds of files are produced; one that is used to form subgroups, one with linkage variables and one with the information we want to extract from different, digital sources. These three files are produced separately in a trusted execution environment, the only common element being a randomly generated, information free serial number.

Applied to the transaction project, the id of bank account owners together with register data from Statistics Norway could be used to form statistical groups. In our system the group allocation program operates in a separate Grouping Compartment, leaves no traces and produces a file which only contain group identifications and the generated serial number. In other situations where there are no id-lists, the group allocation may be based on position data that indicates where people live or businesses are located, data on goods bought and sold to determine peoples gender and age or assign business codes, or other kinds of behavioural patterns. These methods may be less accurate, but still sufficient for statistical purposes.

Again, applied to the transaction project, the linkage information used was the time, location and total on grocery receipts. This information is present both in the customer receipts as well as in the bank transactions. Using combinations of these three information elements we were able to link 66 % of the receipts to serial numbers also found in the statistical groups file. In our system this was done in what we call the Linkage Compartment.

With the help of the common serial number we can produce statistical tables in the Result Compartment which contain daily grocery trade patterns for different subgroups.

# Management issues

Details about this data capture and processing system will be described in the paper. The system introduces a new way of collecting data that differ from survey data in several ways. New data sources may be richer and more accurate than survey data. They may potentially include very accurate information about human action patterns but should in many instances be complemented with data about how people perceive what happens. In the paper we shall outline several topics where digital data can either replace or enhance existing sample surveys or provide new statistics that are currently infeasible via surveys.

As we have already stated, the data capture and processing take place in a trusted execution environment. Even if the suggested system provides all the security measures needed, it needs to gain the trust of data owners, stake holders and the public to be fully accepted. Therefore, it is fundamental that the principles of the system are easy to understand and clearly communicated.

As statistical agencies are increasingly focussing on using new, and often big, data for generating official statistics, the interest and relevance of privacy preserving techniques are growing [2]. Often, these new data sources are privately held and subject to commercial interests. We believe that a secure data capture and processing system like the one described in this paper, has the potential of facilitating access to such data, access being the first and essential step towards utilizing new data sources for statistical purposes.

# References

[1] Fyrberg, Johanna et.al. (2019): Proof of concept - transaction records as a new source for Statistics. Interne dokumenter: 2019/14, Statistics Norway.

[2] Big Data UN Global working group, UN Handbook on Privacy-Preserving Computation Techniques (2019), http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf